

---

# Asymptotically Optimal Agents

---

**Tor Lattimore<sup>1</sup> and Marcus Hutter<sup>1,2</sup>**

Research School of Computer Science

<sup>1</sup>Australian National University and <sup>2</sup>ETH Zürich

{tor.lattimore,marcus.hutter}@anu.edu.au

25 July 2011

## Abstract

Artificial general intelligence aims to create agents capable of learning to solve arbitrary interesting problems. We define two versions of asymptotic optimality and prove that no agent can satisfy the strong version while in some cases, depending on discounting, there does exist a non-computable weak asymptotically optimal agent.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Notation and Definitions</b>	<b>3</b>
<b>3</b>	<b>Non-Existence of Asymptotically Optimal Policies</b>	<b>6</b>
<b>4</b>	<b>Existence of Weak Asymptotically Optimal Policies</b>	<b>10</b>
<b>5</b>	<b>Discussion</b>	<b>15</b>
<b>A</b>	<b>Technical Proofs</b>	<b>19</b>
<b>B</b>	<b>Table of Notation</b>	<b>21</b>

## Keywords

Rational agents; sequential decision theory; artificial general intelligence; reinforcement learning; asymptotic optimality; general discounting.

# 1 Introduction

The dream of artificial general intelligence is to create an agent that, starting with no knowledge of its environment, eventually learns to behave optimally. This means it should be able to learn chess just by playing, or Go, or how to drive a car or mow the lawn, or any task we could conceivably be interested in assigning it.

Before considering the existence of universally intelligent agents, we must be precise about what is meant by optimality. If the environment and goal are known, then subject to computation issues, the optimal policy is easy to construct using an expectimax search from sequential decision theory [NR03]. However, if the true environment is unknown then the agent will necessarily spend some time exploring, and so cannot immediately play according to the optimal policy. Given a class of environments, we suggest two definitions of asymptotic optimality for an agent.

1. An agent is strongly asymptotically optimal if for every environment in the class it plays optimally in the limit.
2. It is weakly asymptotic optimal if for every environment in the class it plays optimally *on average* in the limit.

The key difference is that a strong asymptotically optimal agent must eventually stop exploring, while a weak asymptotically optimal agent may explore forever, but with decreasing frequency.

In this paper we consider the (non-)existence of weak/strong asymptotically optimal agents in the class of all deterministic computable environments. The restriction to deterministic is for the sake of simplicity and because the results for this case are already sufficiently non-trivial to be interesting. The restriction to computable is more philosophical. The Church-Turing thesis is the unprovable hypothesis that anything that can intuitively be computed can also be computed by a Turing machine. Applying this to physics leads to the strong Church-Turing thesis that the universe is computable (possibly stochastically computable, i.e. computable when given access to an oracle of random noise). Having made these assumptions, the largest interesting class then becomes the class of computable (possibly stochastic) environments.

In [Hut04], Hutter conjectured that his universal Bayesian agent, AIXI, was weakly asymptotically optimal in the class of all computable stochastic environments. Unfortunately this was recently shown to be false in [Ors10], where it is proven that no Bayesian agent (with a static prior) can be weakly asymptotically optimal in this class.<sup>1</sup> The key idea behind Orseau’s proof was to show that AIXI eventually stops exploring. This is somewhat surprising because it is normally assumed that Bayesian agents solve the exploration/exploitation dilemma in a principled way. This result is a bit reminiscent of Bayesian (passive induction)

---

<sup>1</sup>Or even the class of computable deterministic environments.

inconsistency results [DF86a, DF86b], although the details of the failure are very different.

We extend the work of [Ors10], where only Bayesian agents are considered, to show that non-computable weak asymptotically optimal agents do exist in the class of deterministic computable environments for some discount functions (including geometric), but not for others. We also show that no asymptotically optimal agent can be computable, and that for all “reasonable” discount functions there does not exist a strong asymptotically optimal agent.

The weak asymptotically optimal agent we construct is similar to AIXI, but with an exploration component similar to  $\epsilon$ -learning for finite state Markov decision processes or the UCB algorithm for bandits. The key is to explore sufficiently often and deeply to ensure that the environment used for the model is an adequate approximation of the true environment. At the same time, the agent must explore infrequently enough that it actually exploits its knowledge. Whether or not it is possible to get this balance right depends, somewhat surprisingly, on how forward looking the agent is (determined by the discount function). That it is sometimes not possible to explore enough to learn the true environment without damaging even a weak form of asymptotic optimality is surprising and unexpected.

Note that the exploration/exploitation problem is well-understood in the Bandit case [ACBF02, BF85] and for (finite-state stationary) Markov decision processes [SL08]. In these restrictive settings, various satisfactory optimality criteria are available. In this work, we do not make any assumptions like Markov, stationary, ergodicity, or else besides computability of the environment. So far, no satisfactory optimality definition is available for this general case.

## 2 Notation and Definitions

We use similar notation to [Hut04, Ors10] where the agent takes actions and the environment returns an observation/reward pair.

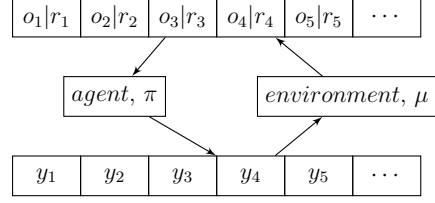
**Strings.** A finite string  $a$  over alphabet  $\mathcal{A}$  is a finite sequence  $a_1a_2a_3\cdots a_{n-1}a_n$  with  $a_i \in \mathcal{A}$ . An infinite string  $\omega$  over alphabet  $\mathcal{A}$  is an infinite sequence  $\omega_1\omega_2\omega_3\cdots$ .  $\mathcal{A}^n$ ,  $\mathcal{A}^*$  and  $\mathcal{A}^\infty$  are the sets of strings of length  $n$ , strings of finite length, and infinite strings respectively. Let  $x$  be a string (finite or infinite) then substrings are denoted  $x_{s:t} := x_sx_{s+1}\cdots x_{t-1}x_t$  where  $s, t \in \mathbb{N}$  and  $s \leq t$ . Strings may be concatenated. Let  $x, y \in \mathcal{A}^*$  of length  $n$  and  $m$  respectively, and  $\omega \in \mathcal{A}^\infty$ . Then define  $xy := x_1x_2\cdots x_{n-1}x_ny_1y_2\cdots y_{m-1}y_m$  and  $x\omega := x_1x_2\cdots x_{n-1}x_n\omega_1\omega_2\omega_3\cdots$ . Some useful shorthands,

$$x_{<t} := x_{1:t-1} \qquad yx_{<t} := y_1x_1y_2x_2\cdots y_{t-1}x_{t-1}. \quad (1)$$

The second of these is ambiguous with concatenation, so wherever  $yx_{<t}$  appears we assume the interleaving definition of (1) is intended. For example, it will be common to see  $yx_{<t}y_t$ , which represents the string  $y_1x_1y_2x_2y_3x_3\cdots y_{t-1}x_{t-1}y_t$ . For

binary strings, we write  $\#1(a)$  and  $\#0(a)$  to mean the number of 0's and number of 1's in  $a$  respectively.

**Environments and Optimality.** Let  $\mathcal{Y}$ ,  $\mathcal{O}$  and  $\mathcal{R} \subset \mathbb{R}$  be action, observation and reward spaces respectively. Let  $\mathcal{X} = \mathcal{O} \times \mathcal{R}$ . An agent interacts with an environment as illustrated in the diagram on the right. First, the agent takes an action, upon which it receives a new observation/reward pair.



The agent then takes another action, receives another observation/reward pair, and so-on indefinitely. The goal of the agent is to maximise its discounted rewards over time. In this paper we consider only deterministic environments where the next observation/reward pair is determined by a function of the previous actions, observations and rewards.

**Definition 1** (Deterministic Environment). A *deterministic environment*  $\mu$  is a function  $\mu : (\mathcal{Y} \times \mathcal{X})^* \times \mathcal{Y} \rightarrow \mathcal{X}$  where  $\mu(yx_{<t}y_t) \in \mathcal{X}$  is the observation/reward pair given after action  $y_t$  is taken in history  $yx_{<t}$ . Wherever we write  $x_t$  we implicitly assume  $x_t = (o_t, r_t)$  and refer to  $o_t$  and  $r_t$  without defining them. An environment  $\mu$  is computable if there exists a Turing machine that computes it.

Note that since environments are deterministic the next observation need not depend on the previous observations (only actions). We choose to leave the dependence as the proofs become clearer when both the action and observation sequence is more visible.

**Assumption 2.**  $\mathcal{Y}$  and  $\mathcal{O}$  are finite,  $\mathcal{R} = [0, 1]$ .

**Definition 3** (Policy). A *policy*  $\pi$  is a function from a history to an action  $\pi : (\mathcal{Y} \times \mathcal{X})^* \rightarrow \mathcal{Y}$ .

As expected, a policy  $\pi$  and environment  $\mu$  can interact with each other to generate a play-out sequence of action/reward/observation tuples.

**Definition 4** (Play-out Sequence). We define the *play-out sequence*  $yx^{\mu, \pi} \in (\mathcal{Y} \times \mathcal{X})^\infty$  inductively by  $y_k^{\mu, \pi} := \pi(yx_{<k}^{\mu, \pi})$  and  $x_k^{\mu, \pi} := \mu(yx_{<k}^{\mu, \pi}y_k^{\mu, \pi})$ .

We need to define the value of a policy  $\pi$  in environment  $\mu$ . To avoid the possibility of infinite rewards, we will use discounted values. While it is common to use only geometric discounting, we have two reasons to allow arbitrary time-consistent discount functions.

1. Geometric discounting has a constant effective horizon, but we feel agents should be allowed to use a discount function that leads to a growing horizon. This is seen in other agents, such as humans, who generally become less myopic as they grow older. See [FOO02] for a overview of generic discounting.

2. The existence of asymptotically optimal agents depends critically on the effective horizon of the discount function.

**Definition 5** (Discount Function). A regular discount function  $\gamma \in \mathbb{R}^\infty$  is a vector satisfying  $\gamma_k \geq 0$  and  $0 < \sum_{t=k}^\infty \gamma_t < \infty$  for all  $k \in \mathbb{N}$ .

The first condition is natural for any definition of a discount function. The second condition is often cited as the purpose of a discount function (to prevent infinite utilities), but economists sometimes use non-summable discount functions, such as hyperbolic. The second condition also guarantees the agent cares about the infinite future, and is required to make asymptotic analysis interesting. We only consider discount functions satisfying all three conditions. In the following, let

$$\Gamma_t := \sum_{i=t}^\infty \gamma_i \quad H_t(p) := \min_{h \in \mathbb{N}} \left\{ h : \frac{1}{\Gamma_t} \sum_{k=t}^{t+h} \gamma_k > p \right\}.$$

An infinite sequence of rewards starting at time  $t$ ,  $r_t, r_{t+1}, r_{t+2}, \dots$  is given a value of  $\frac{1}{\Gamma_t} \sum_{i=t}^\infty \gamma_i r_i$ . The term  $\frac{1}{\Gamma_t}$  is a normalisation term to ensure that values scale in such a way that they can still be compared in the limit. A discount function is computable if there exists a Turing machine computing it. All well known discount functions, such as geometric, fixed horizon and hyperbolic are computable. Note that  $H_t(p)$  exists for all  $p \in [0, 1)$  and represents the effective horizon of the agent. After  $H_t(p)$  time-steps into the future, starting at time  $t$ , the agent stands to gain/lose at most  $1 - p$ .

**Definition 6** (Values and Optimal Policy). The value of policy  $\pi$  when starting from history  $\mathbf{y}_{<t}^{\mu, \pi}$  in environment  $\mu$  is  $V_\mu^\pi(\mathbf{y}_{<t}^{\mu, \pi}) := \frac{1}{\Gamma_t} \sum_{k=t}^\infty \gamma_k r_k^{\mu, \pi}$ . The optimal policy  $\pi_\mu^*$  and its value  $V_\mu^*$  are defined  $\pi_\mu^*(\mathbf{y}_{<t}) := \arg \max_\pi V_\mu^\pi(\mathbf{y}_{<t})$  and  $V_\mu^*(\mathbf{y}_{<t}) := V_\mu^{\pi_\mu^*}(\mathbf{y}_{<t})$ .

Assumption 2 combined with Theorem 6 in [LH11] guarantees the existence of  $\pi_\mu^*$ . Note that the normalisation term  $\frac{1}{\Gamma_t}$  does not change the policy, but is used to ensure that values scale appropriately in the limit. For example, when discounting geometrically we have,  $\gamma_t = \gamma^t$  for some  $\gamma \in (0, 1)$  and so  $\Gamma_t = \frac{\gamma^t}{1-\gamma}$  and  $V_\mu^\pi(\mathbf{y}_{<t}^{\mu, \pi}) = (1 - \gamma) \sum_{k=t}^\infty \gamma^{k-t} r_k^{\mu, \pi}$ .

**Definition 7** (Asymptotic Optimality). Let  $\mathcal{M} = \{\mu_0, \mu_1, \dots\}$  be a finite or countable set of environments and  $\gamma$  be a discount function. A policy  $\pi$  is a *strong asymptotically optimal* policy in  $(\mathcal{M}, \gamma)$  if

$$\lim_{n \rightarrow \infty} [V_\mu^*(\mathbf{y}_{<n}^{\mu, \pi}) - V_\mu^\pi(\mathbf{y}_{<n}^{\mu, \pi})] = 0, \text{ for all } \mu \in \mathcal{M}. \quad (2)$$

It is a *weak asymptotically optimal* policy if

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n [V_\mu^*(\mathbf{y}_{<t}^{\mu, \pi}) - V_\mu^\pi(\mathbf{y}_{<t}^{\mu, \pi})] = 0, \text{ for all } \mu \in \mathcal{M}. \quad (3)$$

Strong asymptotic optimality demands that the value of a *single* policy  $\pi$  converges to the value of the optimal policy  $\pi_\mu^*$  for *all*  $\mu$  in the class. This means that in the limit, a strong asymptotically optimal policy will obtain the maximum value possible in that environments.

Weak asymptotic optimality is similar, but only requires the *average* value of the policy  $\pi$  to converge to the average value of the optimal policy. This means that a weak asymptotically optimal policy can still make infinitely many bad mistakes, but must do so for only a fraction of the time that converges to zero. Strong asymptotic optimality implies weak asymptotic optimality.

While the definition of strong asymptotic optimality is rather natural, the definition of weak asymptotic optimality appears somewhat more arbitrary. The purpose of the average is to allow the agent to make a vanishing fraction of serious errors over its (infinite) life-time. We believe this is a necessary condition for an agent to learn the true environment. Of course, it would be possible to insist that the agent make only  $o(\log n)$  serious errors rather than  $o(n)$ , which would make a stronger version of weak asymptotic optimality. Our choice is the weakest notion of optimality of the above form that still makes sense, which turns out to be already too strong for some discount rates.

Note that for both versions of optimality an agent would be considered optimal if it actively undertook a policy that led it to an extremely bad “hell” state from which it could not escape. Since the state cannot be escaped, its policy would then coincide with the optimal policy and so it would be considered optimal. Unfortunately, this problem seems to be an unavoidable consequence of learning algorithms in non-ergodic environments in general, including the currently fashionable PAC algorithms for arbitrary finite Markov decision processes.

### 3 Non-Existence of Asymptotically Optimal Policies

We present the negative theorem in three parts. The first shows that, at least for computable discount functions, there does not exist a strong asymptotically optimal policy. The second shows that any weak asymptotically optimal policy must be incomputable while the third shows that there exist discount functions for which even incomputable weak asymptotically optimal policies do not exist.

**Theorem 8.** *Let  $\mathcal{M}$  be the class of all deterministic computable environments and  $\gamma$  a computable discount function, then:*

1. *There does not exist a strong asymptotically optimal policy in  $(\mathcal{M}, \gamma)$ .*
2. *There does not exist a computable weak asymptotically optimal policy in  $(\mathcal{M}, \gamma)$ .*

3. If  $\gamma_k := \frac{1}{k(k+1)}$  then there does not exist a weak asymptotically optimal policy in  $(\mathcal{M}, \gamma)$ .

Part 1 of Theorem 8 says there is no strong asymptotically optimal policy in the class of all computable deterministic environments when the discount function is computable. It is likely there exist non-computable discount functions for which there are strong asymptotically optimal policies. Unfortunately the discount functions for which this is true are likely to be somewhat pathological and not realistic.

Given that strong asymptotic optimality is too strong, we should search for weak asymptotically optimal policies. Part 2 of Theorem 8 shows that any such policy is necessarily incomputable. This result features no real new ideas and relies on the fact that you can use a computable policy to hand-craft a computable environment in which it does very badly [Leg06]. In general this approach fails for incomputable policies because the hand-crafted environment will then not be computable. Note that this does not rule out the existence of a stochastically computable weak asymptotically optimal policy.

It turns out that even weak asymptotic optimality is too strong for some discount functions. Part 3 of Theorem 8 gives an example discount function for which no such policy (computable or otherwise) exists. In the next section we introduce a weak asymptotically optimal policy for geometric (and may be extended to other) discounting. Note that  $\gamma_k = \frac{1}{k(k+1)}$  is an example of a discount function where  $H_t(p) = \Omega(t)$ . It is also analytically easy to work with.

All negative results are proven by contradiction, and follow the same basic form.

1. Assume  $\pi$  is a computable/arbitrary weak/strong asymptotically optimal.
2. Therefore  $\pi$  is weak/strong asymptotically optimal in  $\mu$  for some particular  $\mu$ .
3. Construct  $\nu$ , which is indistinguishable from  $\mu$  under  $\pi$ , but where  $\pi$  is not weak/strong asymptotically optimal in  $\nu$ .

*Proof of Theorem 8, Part 1.* Let  $\mathcal{Y} = \{up, down\}$  and  $\mathcal{O} = \emptyset$ . Now assume some policy  $\pi$  is a strong asymptotically optimal policy. Define an environment  $\mu$  by,

$$\mu(y_{<t}y_t) = \begin{cases} \frac{1}{2} & \text{if } y_t = up \\ 0 & \text{if } y_t = down \end{cases} \in \mathcal{R}$$

That is  $\mu(y_{<t}y_t) \in \mathcal{R}$  is the reward given when taking action  $y_t$  having previously taken actions  $y_{<t}$ . Note that we have omitted the observations as  $\mathcal{O} = \emptyset$ . It is easy to see that the optimal policy  $\pi_\mu^*(y_{<t}) = up$  for all  $y_{<t}$  with corresponding value  $V_\mu^*(y_{<t}) = \frac{1}{2}$ . Since  $\pi$  is strongly asymptotically optimal,

$$\lim_{n \rightarrow \infty} V_\mu^\pi(y_{1:n}^{\mu, \pi}) = \frac{1}{2}. \quad (4)$$

Assume there exists a time-sequence  $t_1, t_2, t_3, \dots$  such that  $y_t^{\mu, \pi} = \text{down}$  (and hence  $r_t^{\mu, \pi} = 0$ ) for all  $t \in \bigcup_{i=1}^{\infty} [t_i, t_i + H_{t_i}(\frac{1}{4})]$ . Therefore by the definition of the value function,

$$V_{\mu}^{\pi}(y_{t_i}^{\mu, \pi}) \leq \frac{1}{\Gamma_{t_i}} \left[ \frac{1}{2} \sum_{k=t_i+H_{t_i}(\frac{1}{4})+1}^{\infty} \gamma_k \right] = \frac{1}{2} \left[ 1 - \frac{1}{\Gamma_{t_i}} \sum_{k=t_i}^{t_i+H_{t_i}(\frac{1}{4})} \gamma_k \right] \quad (5)$$

$$\leq \frac{1}{2} \left[ 1 - \frac{1}{4} \right] \quad (6)$$

where (5) follows from the definitions of the value function and  $\Gamma$ , and the assumption in the previous line. (6) follows by algebra and the definition of  $H_{t_i}(\frac{1}{4})$ . This contradicts (4). Therefore for any strong asymptotically optimal policy  $\pi$  there exists a  $T \in \mathbb{N}$  such that for all  $t \geq T$ ,  $y_s^{\mu, \pi} = \text{up}$  for some  $s \in [t, t + H_t(\frac{1}{4})]$ . I.e,  $\pi$  cannot take sub-optimal action *down* too frequently. In particular, it cannot take action *down* for large contiguous blocks of time. Construct a new environment  $\nu$  defined by

$$\nu(y_{<t}y_t) = \begin{cases} \mu(y_{<t}y_t) & \text{if } t < T \\ \frac{1}{2} & \text{if } y_t = \text{up} \\ 1 & \text{if } y_t = \text{down and exist } t' \geq T \text{ such that } t' + H_{t'}(\frac{1}{4}) \leq t \text{ and} \\ & y_s = \text{down } \forall s \in [t', t' + H_{t'}(\frac{1}{4})] \\ 0 & \text{otherwise} \end{cases}$$

Note that  $\nu$  is computable if  $H_t(\frac{1}{4})$  is and that by construction the play-out sequences for  $\mu$  and  $\nu$  when using policy  $\pi$  are identical. We now consider the optimal policy in  $\nu$ . For any  $t \geq T$  consider the value of policy  $\tilde{\pi}$  defined by  $\tilde{\pi}(y_{<t}) := \text{down}$  for all  $y_{<t}$ .

$$\begin{aligned} V_{\nu}^{\tilde{\pi}}(y_{<t}) &= \frac{1}{\Gamma_t} \left[ \sum_{k=t+H_t(\frac{1}{4})}^{\infty} \gamma_k \right] \\ &\geq \frac{3}{4}. \end{aligned}$$

This is because  $\tilde{\pi}$  spends  $H_t(\frac{1}{4}) - 1$  time-steps playing *down* and receiving reward 0 before “unlocking” a reward of 1 on all subsequent plays. On the other hand,  $V_{\nu}^{\pi}(y_{<t}^{\nu, \pi}) \leq \frac{1}{2}$  because  $\pi$  can never unlock the reward of 1 because it never plays *down* for a contiguous block of  $H_t(\frac{1}{4})$  time-steps. By the definition of the optimal policy,  $V_{\nu}^{\tilde{\pi}}(y_{<t}) \leq V_{\nu}^*(y_{<t})$ . Therefore

$$V_{\nu}^*(y_{<t}^{\nu, \pi}) - V_{\nu}^{\pi}(y_{<t}^{\nu, \pi}) \geq \frac{1}{4}.$$



Therefore

$$\limsup_{t \rightarrow \infty} [V_\nu^*(y_{<t}^{\nu, \pi}) - V_\nu^\pi(y_{<t}^{\nu, \pi})] \geq \frac{1}{4} \neq 0.$$

Therefore there does not exist an asymptotically optimal policy  $\pi$  in  $(\mathcal{M}, \gamma)$ .  $\square$

*Proof of Theorem 8, Part 2.* Let  $\mathcal{Y} = \{up, down\}$  and  $\mathcal{O} = \emptyset$ . Now let  $\mathcal{M}$  be the class of all computable deterministic environments and  $\gamma$  be an arbitrary discount function. Suppose  $\pi$  is computable and consider the environment  $\mu$  defined by

$$\mu(y_{<t} y_t) = \begin{cases} 1 & \text{if } y_t \neq \pi(y_{>t}) \\ 0 & \text{otherwise} \end{cases}$$

Since  $\pi$  is computable  $\mu$  is as well. Therefore  $\mu \in \mathcal{M}$ . Now  $V_\mu^*(y_{<t}) = 1$  for all  $y_{<t}$  while  $V_\mu^\pi(y_{<t}) = 0$ . Therefore  $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n |V_\mu^*(y_{<t}) - V_\mu^\pi(y_{<t})| = 1$  and so  $\pi$  is not weakly asymptotically optimal.  $\square$

*Proof of Theorem 8, Part 3.* Recall  $\gamma_k = \frac{1}{k(k+1)}$  and so  $\Gamma_t = \frac{1}{t}$ . Now let  $\mathcal{Y} = \{up, down\}$  and  $\mathcal{O} = \emptyset$ . Define  $\mu$  by

$$\mu(y_{<t} y_t) = \begin{cases} \frac{1}{2} & \text{if } y_t = up \\ \frac{1}{2} - \epsilon & \text{if } y_t = down \end{cases}$$

where  $\epsilon \in (0, \frac{1}{2})$  will be chosen later. As before,  $V_\mu^*(y_{<t}) = \frac{1}{2}$ . Assume  $\pi$  is weakly asymptotically optimal. Therefore

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n V_\mu^\pi(y_{<t}^{\mu, \pi}) = \frac{1}{2}. \quad (7)$$

We show by contradiction that  $\pi$  cannot explore (take action *down*) too often. Assume there exists an infinite time-sequence  $t_1, t_2, t_3, \dots$  such that  $\pi(y_{<t}^{\mu, \pi}) = down$  for all  $t \in \bigcup_{i=1}^\infty [t_i, 2t_i]$ . Then for  $t \in [t_i, \frac{3}{2}t_i]$  we have

$$V_\mu^\pi(y_{<t}^{\mu, \pi}) \equiv \frac{1}{\Gamma_t} \sum_{k=t}^\infty \gamma_k r_k^{\mu, \pi} \leq t \left[ \left( \frac{1}{2} - \epsilon \right) \sum_{k=t}^{2t_i} \gamma_k + \frac{1}{2} \sum_{k=2t_i+1}^\infty \gamma_k \right] \quad (8)$$

$$= \frac{1}{2} - \epsilon \left[ 1 - \frac{t}{2t_i + 1} \right] < \frac{1}{2} - \frac{\epsilon}{4} \quad (9)$$

where (8) is the definition of the value function and the previous assumption and definition of  $\mu$ . (9) by algebra and since  $t \in [t_i, \frac{3}{2}t_i]$ . Therefore

$$\frac{1}{2t_i} \sum_{t=1}^{2t_i} V_\mu^\pi(y_{<t}^{\mu, \pi}) < \frac{1}{2t_i} \left[ \sum_{t=1}^{t_i-1} \frac{1}{2} + \sum_{t=t_i}^{\frac{3}{2}t_i-1} \left( \frac{1}{2} - \frac{\epsilon}{4} \right) + \sum_{t=\frac{3}{2}t_i}^{2t_i} \frac{1}{2} \right] = \frac{1}{2} - \frac{1}{16}\epsilon. \quad (10)$$

The first inequality follows from (9) and because the maximum value of any play-out sequence in  $\mu$  is  $\frac{1}{2}$ . The second by algebra. Therefore  $\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n V_{\mu}^{\pi}(y_{<t}^{\mu, \pi}) < \frac{1}{2} - \frac{1}{16}\epsilon < \frac{1}{2}$ , which contradicts (7). Therefore there does not exist a time-sequence  $t_1 < t_2 < t_3 < \dots$  such that  $\pi(y_{<t}^{\mu, \pi}) = \text{down}$  for all  $t \in \bigcup_{i=1}^{\infty} [t_i, 2t_i]$ .

So far we have shown that  $\pi$  cannot “explore” for  $t$  consecutive time-steps starting at time-step  $t$ , infinitely often. We now construct an environment similar to  $\mu$  where this is required. Choose  $T$  to be larger than the last time-step  $t$  at which  $y_s^{\mu, \pi} = \text{down}$  for all  $s \in [t, 2t]$ . Define  $\nu$  by

$$\nu(y_{<t}y_t) = \begin{cases} \mu(y_{<t}y_t) & \text{if } t < T \\ \frac{1}{2} & \text{if } y_t = \text{down and there does not exist } t' \geq T \\ & \text{such that } y_s = \text{down} \forall s \in [t', 2t'] \\ 1 & \text{if } y_t = \text{down and exists } t' \geq T \text{ such that } 2t' < t \text{ and} \\ & y_s = \text{down} \forall s \in [t', 2t'] \\ \frac{1}{2} - \epsilon & \text{otherwise} \end{cases}$$

Now we compare the values in environment  $\nu$  of  $\pi$  and  $\pi_{\nu}^*$  at times  $t \geq T$ . Since  $\pi$  does not take action *down* for  $t$  consecutive time-steps at any time after  $T$ , it never “unlocks” the reward of 1 and so  $V_{\nu}^{\pi}(y_{<t}^{\nu, \pi}) \leq \frac{1}{2}$ . Now let  $\tilde{\pi}(y_{<t}) = \text{down}$  for all  $y_{<t}$ . Therefore, for  $t \geq 2T$ ,

$$V_{\nu}^{\tilde{\pi}}(y_{<t}^{\nu, \pi}) \equiv \frac{1}{\Gamma_t} \sum_{k=t}^{\infty} \gamma_k r_k^{\nu, \tilde{\pi}} \geq t \left[ \left( \frac{1}{2} - \epsilon \right) \sum_{k=t}^{2t-1} \gamma_k + \sum_{k=2t}^{\infty} \gamma_k \right] \quad (11)$$

$$= t \left[ \left( \frac{1}{2} - \epsilon \right) \left( \frac{1}{t} - \frac{1}{2t} \right) + \frac{1}{2t} \right] = \frac{3}{4} - \frac{1}{2}\epsilon \quad (12)$$

where (11) follows by the definition of  $\nu$  and  $\tilde{\pi}$ . (12) by the definition of  $\gamma_k$  and algebra. Finally, setting  $\epsilon = \frac{1}{4}$  gives  $V_{\nu}^{\tilde{\pi}}(y_{<t}^{\nu, \pi}) \geq \frac{5}{8} = \frac{1}{2} + \frac{1}{8}$ . Since  $V_{\nu}^* \geq V_{\nu}^{\tilde{\pi}}$ , we get  $V_{\nu}^*(y_{<t}^{\nu, \pi}) - V_{\nu}^{\pi}(y_{<t}^{\nu, \pi}) \geq V_{\nu}^{\tilde{\pi}}(y_{<t}^{\nu, \pi}) - V_{\nu}^{\pi}(y_{<t}^{\nu, \pi}) \geq \frac{1}{8}$ . Therefore  $\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n [V^*(y_{<t}^{\nu, \pi}) - V_{\nu}^{\pi}(y_{<t}^{\nu, \pi})] \geq \frac{1}{8}$ , and so  $\pi$  is not weakly asymptotically optimal.  $\square$

We believe it should be possible to generalise the above to computable discount functions with  $H_t(p) > c_p t$  with  $c_p > 0$  for infinitely many  $t$ , but the proof will likely be messy.

## 4 Existence of Weak Asymptotically Optimal Policies

In the previous section we showed there did not exist a strong asymptotically optimal policy (for most discount functions) and that any weak asymptotically optimal policy

must be incomputable. In this section we show that a weak asymptotically optimal policy exists for geometric discounting (and is, of course, incomputable).

The policy is reminiscent of  $\epsilon$ -exploration in finite state MDPs (or UCB for bandits) in that it spends most of its time exploiting the information it already knows, while still exploring sufficiently often (and for sufficiently long) to detect any significant errors in its model.

The idea will be to use a model-based policy that chooses its current model to be the first environment in the model class (all computable deterministic environments) consistent with the history seen so far. With increasing probability it takes the best action according to this policy, while still occasionally exploring randomly. When it explores it always does so in bursts of increasing length.

**Definition 9** (History Consistent). A deterministic environment  $\mu$  is *consistent* with history  $y_{<t}$  if  $\mu(y_{<k}y_k) = x_k$ , for all  $k < t$ .

**Definition 10** (Weak Asymptotically Optimal Policy). Let  $\mathcal{Y} = \{0, 1\}$  and  $\mathcal{M} = \{\mu_1, \mu_2, \mu_3, \dots\}$  be a countable class of deterministic environments. Define a probability measure  $P$  on  $\mathcal{B}^\infty$  inductively by,  $P(z_n = 1 | z_{<n}) := \frac{1}{n}$ , for all  $z_{<n} \in \mathcal{B}^{n-1}$ . Now let  $\chi \in \mathcal{B}^\infty$  be sampled from  $P$  and define  $\bar{\chi}, \dot{\chi}^h \in \mathcal{B}^\infty$  by

$$\bar{\chi}_k := \begin{cases} 1 & \text{if } k \in \bigcup_{i:\chi_i=1} [i, i + \log i] \\ 0 & \text{otherwise} \end{cases} \quad \dot{\chi}_k^h := \begin{cases} 0 & \text{if } \bar{\chi}_{k:k+h} = 0^{h+1} \\ 1 & \text{otherwise} \end{cases}$$

Next let  $\psi$  be sampled from the uniform measure (each bit of  $\psi$  is independently sampled from a Bernoulli 1/2 distribution) and define a policy  $\pi$  by,

$$\pi(y_{<t}) := \begin{cases} \pi_{\nu_t}^*(y_{<t}^{\pi, \mu}) & \text{if } \bar{\chi}_t = 0 \\ \psi_t & \text{otherwise} \end{cases} \quad (13)$$

where  $\nu_t = \mu_{i_t}$  with  $i_t = \min \{i : \mu_i \text{ consistent with history } y_{<t}^{\pi, \mu}\} < \infty$ . Note that  $i_t$  is always finite because there exists an  $i$  such that  $\mu_i = \mu$ , in which case  $\mu_i$  is necessarily consistent with  $y_{<t}^{\pi, \mu}$ .

Intuitively,  $\chi_k = 1$  at time-steps when the agent will explore for  $\log k$  time-steps.  $\bar{\chi}_k = 1$  if the agent is exploring at time  $k$  and  $\psi_k$  is the action taken if exploring at time-step  $k$ .  $\dot{\chi}$  will be used later, with  $\dot{\chi}_k^h = 1$  if the agent will explore at least once in the interval  $[k, k + h]$ . If the agent is not exploring then it acts according to the optimal policy for the first consistent environment in  $\mathcal{M}$ .

**Theorem 11.** Let  $\gamma_t = \gamma^t$  with  $\gamma \in (0, 1)$  (geometric discounting) then the policy defined in Definition 10 is weakly asymptotically optimal in the class of all deterministic computable environments with probability 1.

Some remarks:

1. That  $\mathcal{Y} = \{0, 1\}$  is only convenience, rather than necessity. The policy is easily generalised to arbitrary finite  $\mathcal{Y}$ .
2.  $\pi$  is essentially a stochastic policy. With some technical difficulties it is possible to construct an equivalent deterministic policy. This is done by choosing  $\chi$  to be any  $P$ -Martin-Löf random sequence and  $\psi$  to be a sequence that is Martin-Löf random w.r.t to the uniform measure. The theorem then holds for *all* deterministic environments. The proof is somewhat delicate and may not extend nicely to stochastic environments. For an introduction to Kolmogorov complexity and Martin-Löf randomness, see [LV08]. For a reason why the stochastic case may not go through as easily, see [HM07].
3. The policy defined in Definition 10 is not computable for two reasons. First, because it relies on the stochastic sequences  $\chi$  and  $\psi$ . Second, because the operation of finding the first environment consistent with the history is not computable.<sup>2</sup> We do not know if there exists a weak asymptotically optimal policy that is computable when given access to a random number generator (or if it is given  $\chi$  and  $\psi$ ).
4. The bursts of exploration are required for optimality. Without them it will be possible to construct counter-example environments similar to those used in part 3 of Theorem 8.

Before the proof we require some more definitions and lemmas. Easier proofs are omitted.

**Definition 12** (*h-Difference*). Let  $\mu$  and  $\nu$  be two environments consistent with history  $y_{x_{<t}}$ , then  $\mu$  is *h-different* to  $\nu$  if there exists  $y_{x_{t:t+h}}$  satisfying

$$\begin{aligned} y_k &= \pi_\mu^*(y_{x_{<k}}) \text{ for all } k \in [t, t+h], \\ x_k &= \mu(y_{x_{<k}}y_k) \text{ for all } k \in [t, t+h], \\ x_k &\neq \nu(y_{x_{<k}}y_k) \text{ for some } k \in [t, t+h]. \end{aligned}$$

Intuitively,  $\mu$  is *h-different* to  $\nu$  at history  $y_{x_{<t}}$  if playing the optimal policy for  $\mu$  for  $h$  time-steps makes  $\nu$  inconsistent with the new history. Note that *h-difference* is *not* symmetric.

**Lemma 13.** If  $a_n \in [0, 1]$  and  $\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n a_i = \epsilon$  and  $\alpha \in \mathcal{B}^\infty$  is an indicator sequence with  $\alpha_i := \llbracket a_i \geq \epsilon/4 \rrbracket$ ,<sup>3</sup> then  $\prod_{i=1}^\infty \left[1 - \frac{\alpha_i}{i}\right] = 0$ .

See the appendix for the proof.

**Lemma 14.** Let  $a_1, a_2, a_3, \dots$  be a sequence with  $a_n \in [0, 1]$  for all  $n$ . The following properties of  $\chi$  are true with probability 1.

---

<sup>2</sup>The class of computable environments is not recursively enumerable [LV08].

<sup>3</sup> $\llbracket \text{expression} \rrbracket = 1$  if *expression* is true and 0 otherwise.

1. For any  $h$ ,  $\limsup_{n \rightarrow \infty} \frac{1}{n} \#1(\dot{\chi}_{1:n}^h) = 0$ .
2. If  $\limsup \frac{1}{n} \sum_{i=1}^n a_i = \epsilon > 0$  and  $\alpha_i := \llbracket a_i > \epsilon/2 \rrbracket$  then  $\alpha_i = \chi_i = 1$  for infinitely many  $i$ .

*Proof.* 1. Let  $i \in \mathbb{N}$ ,  $\epsilon > 0$  and  $E_i^\epsilon$  be the event that  $\#1(\dot{\chi}_{1:2^i}^h) > 2^i \epsilon$ . Using the definition of  $\dot{\chi}^h$  to compute the expectation  $\mathbf{E}[\#1(\dot{\chi}_{1:2^i}^h)] < i(i+1)h$  and applying the Markov inequality gives that  $P(E_i^\epsilon) < i(i+1)h2^{-i}/\epsilon$ . Therefore  $\sum_{i \in \mathbb{N}} P(E_i^\epsilon) < \infty$ . Therefore the Borel-Cantelli lemma gives that  $E_i^\epsilon$  occurs for only finitely many  $i$  with probability 1. We now assume that  $\limsup_{n \rightarrow \infty} \frac{1}{n} \#1(\dot{\chi}_{1:n}^h) > 2\epsilon > 0$  and show that  $E_i^\epsilon$  must occur infinitely often. By the definition of  $\limsup$  and our assumption we have that there exists a sequence  $n_1, n_2, \dots$  such that  $\#1(\dot{\chi}_{1:n_i}^h) > 2n_i \epsilon$  for all  $i \in \mathbb{N}$ . Let  $n^+ := \min_{k \in \mathbb{N}} \{2^k : 2^k \geq n\}$  and note that  $\#1(\dot{\chi}_{1:n_i^+}^h) > n_i^+ \epsilon$ , which is exactly  $E_{\log n_i^+}^\epsilon$ . Therefore there exist infinitely many  $i$  such that  $E_i^\epsilon$  occurs and so  $\limsup_{n \rightarrow \infty} \frac{1}{n} \#1(\dot{\chi}_{1:n}^h) = 0$  with probability 1.

2. The probability that  $\alpha_i = 1 \implies \chi_i = 0$  for all  $i \geq T$  is  $P(\alpha_i = 1 \implies \chi_i = 0 \forall i \geq T) = \prod_{i=T}^\infty (1 - \frac{\alpha_i}{i}) =: p = 0$ , by Lemma 13. Therefore the probability that  $\alpha_i = \chi_i = 1$  for only finitely many  $i$  is zero. Therefore there exists infinitely many  $i$  with  $\alpha_i = \chi_i = 1$  with probability 1, as required.  $\square$

**Lemma 15** (Approximation Lemma). *Let  $\pi_1$  and  $\pi_2$  be policies,  $\mu$  an environment and  $h \geq H_t(1 - \epsilon)$ . Let  $\mathbf{y}_{<t}$  be an arbitrary history and  $\mathbf{y}_{t:t+h}^{\mu, \pi_i}$  be the future action/observation/reward triples when playing policy  $\pi_i$ . If  $\mathbf{y}_{t:t+h}^{\pi_1, \mu} = \mathbf{y}_{t:t+h}^{\pi_2, \mu}$  then  $|V_\mu^{\pi_1}(\mathbf{y}_{<t}) - V_\mu^{\pi_2}(\mathbf{y}_{<t})| < \epsilon$ .*

*Proof.* By the definition of the value function,

$$|V_\mu^{\pi_1}(\mathbf{y}_{<t}) - V_\mu^{\pi_2}(\mathbf{y}_{<t})| \leq \frac{1}{\Gamma_t} \sum_{i=t}^\infty \gamma_i |r_i^{\pi_1, \mu} - r_i^{\pi_2, \mu}| \quad (14)$$

$$= \frac{1}{\Gamma_t} \sum_{i=t+h+1}^\infty \gamma_i |r_i^{\pi_1, \mu} - r_i^{\pi_2, \mu}| \leq \frac{1}{\Gamma_t} \sum_{i=t+h+1}^\infty \gamma_i < \epsilon \quad (15)$$

(14) follows from the definition of the value function. (15) since  $r_i^{\pi_1, \mu} = r_i^{\pi_2, \mu}$  for  $i \in [t, t+h]$ , rewards are bounded in  $[0, 1]$  and by the definition of  $h := H_t(1 - \epsilon)$  (Definition 5).  $\square$

Recall that  $\pi_\mu^*$  and  $\pi_\nu^*$  are the optimal policies in environments  $\mu$  and  $\nu$  respectively (see Definition 6).

**Lemma 16** ( $h$ -difference). *If  $|V_\mu^{\pi_\mu^*}(\mathbf{y}_{<t}^{\pi, \mu}) - V_\mu^{\pi_\nu^*}(\mathbf{y}_{<t}^{\pi, \mu})| > \epsilon$  then  $\mu$  is  $H_t(1 - \epsilon)$ -different to  $\nu$  on  $\mathbf{y}^{\pi, \mu}$ .*

*Proof.* Follows from the approximation lemma.  $\square$

We are now ready to prove the main theorem.

*Proof of Theorem 11.* Let  $\pi$  be the policy defined in Definition 10 and  $\mu$  be the true (unknown) environment. Recall that  $\nu_t = \mu_{i_t}$  with  $i_t = \min \{i : \mu_i \text{ consistent with history } \mathcal{H}_{< t}^{\pi, \mu}\}$  is the first model consistent with the history  $\mathcal{H}_{< t}^{\pi, \mu}$  at time  $t$  and is used by  $\pi$  when not exploring. First we claim there exists a  $T \in \mathbb{N}$  and environment  $\nu$  such that  $\nu_t = \nu$  for all  $t \geq T$ . Two facts,

1. If  $\mu_i$  is inconsistent with history  $\mathcal{H}_{< t}^{\pi, \mu}$  then it is also inconsistent with  $\mathcal{H}_{< t+h}^{\pi, \mu}$  for all  $h \in \mathbb{N}$ .
2.  $\mu$  is consistent with  $\mathcal{H}_{< t}^{\pi, \mu}$  for all  $\pi, t$ .

By 1) we have that the sequence  $i_1, i_2, i_3, \dots$  is monotone increasing. By 2) we have that the sequence is bounded by  $i$  with  $\mu_i = \mu$ . The claim follows since any bounded monotone sequence of natural numbers converges in finite time. Let  $\nu := \nu_\infty$  be the environment to which  $\nu_1, \nu_2, \nu_3, \dots$  converges to. Note that  $\nu$  must be consistent with history  $\mathcal{H}_{< t}^{\mu, \pi}$  for all  $t$ . We now show by contradiction that the optimal policy for  $\nu$  is weakly asymptotically optimal in environment  $\mu$ . Suppose it were not, then

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n [V_\mu^*(\mathcal{H}_{< t}^{\pi, \mu}) - V_\mu^{\pi_\nu}(\mathcal{H}_{< t}^{\pi, \mu})] = \epsilon > 0. \quad (16)$$

Let  $\alpha \in \mathcal{B}^\infty$  be defined by  $\alpha_t := 1$  if and only if,

$$[V_\mu^*(\mathcal{H}_{< t}^{\pi, \mu}) - V_\mu^\pi(\mathcal{H}_{< t}^{\pi, \mu})] \geq \epsilon/4. \quad (17)$$

By Lemma 14 there exists (with probability one) an infinite sequence  $t_1, t_2, t_3, \dots$  for which  $\chi_k = \alpha_k = 1$ . Intuitively we should view time-step  $t_k$  as the start of an “exploration” phase where the agent explores for  $\log t_k$  time-steps. Let  $h := H_{t_k}(1 - \epsilon/4) = \lceil \log(\epsilon/4) / \log \gamma \rceil$ , which importantly is independent of  $t_k$  (for geometric discounting). Since  $\log t_k \rightarrow \infty$  we will assume that  $\log t_k \geq h$  for all  $t_k$ . Therefore  $\bar{\chi}_i = 1$  for all  $i \in \bigcup_{k=1}^\infty [t_k, t_k + h]$ . Therefore by the definition of  $\pi$ ,  $\pi(\mathcal{H}_{< i}^{\pi, \mu}) = \psi_i$  for  $i \in \bigcup_{k=1}^\infty [t_k, t_k + h]$ . By Lemma 16 and Equation (17),  $\mu$  is  $h$ -different to  $\nu$  on history  $\mathcal{H}_{< t_k}^{\pi, \mu}$ . This means that if there exists a  $k$  such that  $\pi$  plays according to the optimal policy for  $\mu$  on all time-steps  $t \in [t_k, t_k + h]$  then  $\nu$  will be inconsistent with the history  $\mathcal{H}_{1:t_k+h}^{\mu, \pi}$ , which is a contradiction. We now show that  $\pi$  does indeed play according to the optimal policy for  $\mu$  for all time-steps  $t \in [t_k, t_k + h]$  for at least one  $k$ . Formally, we show the following holds with probability 1 for some  $k$ .

$$\psi_i \equiv \pi(\mathcal{H}_{< i}^{\pi, \mu}) = \pi_\mu^*(\mathcal{H}_{< i}^{\pi, \mu}), \text{ for all } i \in [t_k, t_k + h]. \quad (18)$$

Recall that  $\psi \in \mathcal{B}^\infty$  where  $\psi_i \in \mathcal{B}$  is identically independently distributed according to a Bernoulli( $\frac{1}{2}$ ) distribution. Therefore  $P(\psi_i = \pi_\mu^*(\mathcal{H}_{< i}^{\pi, \mu})) = \frac{1}{2}$ . Therefore  $p := P(\psi_i = \pi_\mu^*(\mathcal{H}_{< i}^{\pi, \mu}) \forall i \in [t_k, t_k + h]) = \prod_{i=t_k}^{t_k+h} P(\psi_i = \pi_\mu^*(\mathcal{H}_{< i}^{\pi, \mu})) = 2^{-h-1} > 0$  and  $P(\forall k \exists i \in [t_k, t_k + h] \text{ with } \psi_i \neq \pi_\mu^*(\mathcal{H}_{< i}^{\pi, \mu})) = \prod_{k=1}^\infty (1 - p) = 0$ . Therefore Equation

(18) is satisfied for some  $k$  with probability 1 and so Equation (16) leads to a contradiction. Therefore

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n [V_{\mu}^*(\mathcal{Y}_{<t}^{\pi, \mu}) - V_{\mu}^{\pi^*}(\mathcal{Y}_{<t}^{\pi, \mu})] = 0. \quad (19)$$

We have shown that the optimal policy for  $\nu$  has similar  $\mu$ -values to the optimal policy for  $\mu$ . We now show that  $\pi$  acts according to  $\pi_{\nu}^*$  sufficiently often that it too has values close to those of the optimum policy for the true environment,  $\mu$ . Let  $\epsilon > 0$ ,  $h := H_t(1 - \epsilon)$  and  $t \geq T$ . If  $\dot{\chi}_t^h = 0$  then by the definition of  $\pi$  and the approximation lemma we obtain

$$|V_{\mu}^{\pi^*}(\mathcal{Y}_{<t}^{\pi, \mu}) - V_{\mu}^{\pi}(\mathcal{Y}_{<t}^{\pi, \mu})| < \epsilon. \quad (20)$$

Therefore

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n |V_{\mu}^{\pi^*}(\mathcal{Y}_{<t}^{\pi, \mu}) - V_{\mu}^{\pi}(\mathcal{Y}_{<t}^{\pi, \mu})| \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \left| \sum_{t=1}^{T-1} 1 + \sum_{t=T}^n [\dot{\chi}_t^h(1 - \epsilon) + \epsilon] \right| \quad (21)$$

$$= \epsilon + (1 - \epsilon) \limsup_{n \rightarrow \infty} \frac{1}{n} \#1(\dot{\chi}_{T:n}^h) \quad (22)$$

$$= \epsilon \quad (23)$$

where (21) follows since values are bounded in  $[0, 1]$  and Equation (20). (22) follows by algebra. (23) by part 1 of Lemma 14. By sending  $\epsilon \rightarrow 0$ ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n [V_{\mu}^{\pi^*}(\mathcal{Y}_{<t}^{\pi, \mu}) - V_{\mu}^{\pi}(\mathcal{Y}_{<t}^{\pi, \mu})] = 0. \quad (24)$$

Finally, combining Equations (19) and (24) gives the result.  $\square$

We expect this theorem to generalise without great difficulty to discount functions satisfying  $H_t(p) < c_p \log(t)$  for all  $p$ . There will be two key changes. First, extend the exploration time to some function  $E(t)$  with  $E(t) \in O(H_p(t))$  for all  $p$ . Second, modify the probability of exploration to ensure that Lemma 14 remains true.

## 5 Discussion

**Summary.** Part 1 of Theorem 8 shows that no policy can be strongly asymptotically optimal for any computable discount function. The key insight is that strong asymptotic optimality essentially implies exploration must eventually cease. Once this occurs, the environment can change without the agent discovering the difference and the policy will no longer be optimal.

A weaker notion of asymptotic optimality, that a policy be optimal on average in the limit, turns out to be more interesting. Part 2 of Theorem 8 shows that no weak asymptotically optimal policy can be computable. We should not be surprised by this result. Any computable policy can be used to construct a computable environment in which that policy does very badly. Note that by computable here we mean deterministic and computable. There may be computable stochastic policies that are weakly asymptotically optimal, but we feel this is unlikely.

Part 3 of Theorem 8, shows that even weak asymptotically optimal policies need not exist if the discount function is sufficiently far-sighted. On the other hand, Theorem 11 shows that weak asymptotically optimal policies do exist for some discount rates, in particular, for the default geometric discounting. These non-trivial and slightly surprising result shows that choice of discount function is crucial to the existence of weak asymptotically optimal policies. Where weak asymptotically optimal policies do exist, they must explore infinitely often and in increasing contiguous bursts of exploration where the length of each burst is dependent on the discount function.

**Consequences.** It would appear that Theorem 8 is problematic for artificial general intelligence. We cannot construct incomputable policies, and so we cannot construct weak asymptotically optimal policies. However, this is not as problematic as it may seem. There are a number of reasonable counter arguments:

1. We may be able to make stochastically computable policies that are asymptotically optimal. If the existence of true random noise is assumed then this would be a good solution.
2. The counter-example environment constructed in part 2 of Theorem 8 is a single environment roughly as complex as the policy itself. Certainly, if the world were adversarial this would be a problem, but in general this appears not to be the case. On the other hand, if the environment is a learning agent itself, this could result in a complexity arms race without bound. There may exist a computable weak asymptotically optimal policy in some extremely large class of environments. For example, the algorithm of Section 4 is stochastically computable when the class of environments is recursively enumerable and contains only computable environments. A natural (and already quite large) class satisfying these properties is finite-state Markov decision processes with  $\{0, 1\}$ -valued transition functions and rational-valued rewards.
3. While it is mathematically pleasant to use asymptotic behaviour to characterise optimal general intelligent behaviour, in practise we usually care about more immediate behaviour. We expect that results, and even (parameter free) formal definitions of intelligence satisfying this need will be challenging, but worthwhile.
4. Accept that even weak asymptotic optimality is too strong and find something weaker, but still useful.



**Relation to AIXI.** The policy defined in Section 4 is not equivalent to AIXI [Hut04], which is also incomputable. However, if the computable environments in  $\mathcal{M}$  are ordered by complexity then it is likely the two will be quite similar. The key difference is the policy defined in this paper will continue to explore whereas it was shown in [Ors10] that AIXI eventually ceases exploration in some environments and some histories. We believe, and a proof should not be too hard, that AIXI will become weakly asymptotically optimal if an exploration component is added similarly as in Section 4.

We now briefly compare the self-optimising property in [Hut02] to strong asymptotic optimality. A policy  $\pi$  is self-optimising in a class  $\mathcal{M}$  if  $\lim_{t \rightarrow \infty} [V_{\mu}^*(y_{x_{<t}}) - V_{\mu}^{\pi}(y_{x_{<t}})] = 0$  for any infinite history  $y_{1:\infty}$  and for all  $\mu \in \mathcal{M}$ . This is similar to strong asymptotic optimality, but convergence must be on all histories, rather than the histories actually generated by  $\pi$ . This makes the self-optimising property a substantially stronger form of optimality than strong asymptotic optimality. It has been proven that if there exists self-optimising policy for a particular class, then AIXI is also self-optimising in that class [Hut02].

It is possible to define a weak version of the self-optimising property by insisting that  $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n [V_{\mu}^*(y_{x_{<t}}) - V_{\mu}^{\pi}(y_{x_{<t}})] = 0$  for all  $y_{1:\infty}$  and all  $\mu \in \mathcal{M}$ . It can then be proven that the existence of a weak self-optimising policy would imply that AIXI were also weakly self-optimising. However, the policy defined in Section 4 cannot be modified to have the weak self-optimising property. It must be allowed to choose its actions itself. This is consistent with the work in [Ors10] which shows that AIXI cannot be weakly asymptotically optimal, and so cannot be weak self-optimising either.

**Discounting.** Throughout this paper we have assumed rewards to be discounted according to a summable discount function. A very natural alternative to discounting, suggested in [LH07], is to restrict interest to environments satisfying  $\sum_{k=1}^{\infty} r_k^{\mu, \pi} \leq 1$ . Now the goal of the agent is simply to maximise summed rewards. In this setting it is easy to see that the positive theorem is lost while all negative ones still hold! This is unfortunate, as discounting presents a major philosophical challenge. How to choose a discount function?

**Assumptions/Limitations.** Assumption 2 ensures that  $\mathcal{Y}$  and  $\mathcal{O}$  are finite. All negative results go through for countable  $\mathcal{Y}$  and  $\mathcal{O}$ . The optimal policy of Section 4 may not generalise to countable  $\mathcal{Y}$ . We have also assumed bounded reward and discrete time. The first seems reasonable while the second allows for substantially easier analysis. Additionally we have only considered deterministic computable environments. The stochastic case is unquestionably interesting. We invoked Church thesis to assert that computable stochastic environments are essentially the largest class of interesting environments.

Many of our Theorems are only applicable to computable discount functions. All well-known discount function in use today are computable. However [Hut04] has suggested  $\gamma_t = 2^{-K(t)}$ , where  $K(t)$  is the (incomputable) prefix Kolmogorov

complexity of  $t$ , may have nice theoretical properties.

**Open Questions.** A number of open questions have arisen during this research. In particular,

1. Prove Theorem 8 for a larger class of discount functions.
2. Prove or disprove the existence of a weak asymptotically optimal stochastically computable policy for some discount function in the class of deterministic computable environments.
3. Modify the policy of Section 4 to the larger class of stochastically computable environments. We believe this to be possible along the lines of [RH08], but inevitably the analysis will be messy and complex.
4. Extend Part 3 of Theorem 8 and Theorem 11 to a complete classification of discount functions according to whether or not they admit a weak asymptotically optimal policy in the class of computable environments.
5. Prove that AIXI is weakly asymptotically optimal when augmented with an exploration component as in Section 4.
6. Define and study other formal measures of optimality/intelligence.

**Acknowledgements.** We thank Laurent Orseau, Wen Shao and reviewers for valuable feedback on earlier drafts and the Australian Research Council for support under grant DP0988049.

## References

- [ACBF02] Peter Auer, Nicoló Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.
- [BF85] Donald A. Berry and Bert Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London, 1985.
- [DF86a] P. Diaconis and D. Freedman. On inconsistent Bayes estimates of location. *The Annals of Statistics*, 14(1):pp. 68–87, 1986.
- [DF86b] Persi Diaconis and David Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):pp. 1–26, 1986.
- [FOO02] Shane Frederick, George L. Oewenstein, and Ted O’Donoghue. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40(2), 2002.
- [HM07] Marcus Hutter and Andrej A. Muchnik. On semimeasures predicting Martin-Löf random sequences. *Theoretical Computer Science*, 382(3):247–261, 2007.

- [Hut02] Marcus Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Proc. 15th Annual Conf. on Computational Learning Theory (COLT'02)*, volume 2375 of *LNAI*, pages 364–379, Sydney, 2002. Springer, Berlin.
- [Hut04] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004.
- [Leg06] Shane Legg. Is there an elegant universal theory of prediction? In Jos Balczar, Philip Long, and Frank Stephan, editors, *Algorithmic Learning Theory*, volume 4264 of *Lecture Notes in Computer Science*, pages 274–287. Springer Berlin / Heidelberg, 2006.
- [LH07] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds & Machines*, 17(4):391–444, 2007.
- [LH11] Tor Lattimore and Marcus Hutter. Time consistent discounting. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011.
- [LV08] Ming Li and Paul Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, Verlag, 3rd edition, 2008.
- [NR03] Peter Norvig and Stuart J. Russell. *Artificial Intelligence: A Modern Approach (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 0002 edition, 2003.
- [Ors10] Laurent Orseau. Optimality issues of universal greedy agents with static priors. In Marcus Hutter, Frank Stephan, Vladimir Vovk, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 6331 of *Lecture Notes in Computer Science*, pages 345–359. Springer Berlin / Heidelberg, 2010.
- [RH08] Daniil Ryabko and Marcus Hutter. On the possibility of learning in reactive environments with arbitrary dependence. *Theoretical Computer Science*, 405(3):274–284, 2008.
- [SL08] Alexander L. Strehl and Michael L. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

## A Technical Proofs

**Lemma 17.** *Let  $A = \{a_1, a_2, \dots, a_n\}$  with  $a \in [0, 1]$  for all  $a \in A$ . If  $\frac{1}{n} \sum_{a \in A} a \geq \epsilon$  then*

$$\left| \left\{ a \in A : a \geq \frac{\epsilon}{2} \right\} \right| > n \frac{\epsilon}{2}$$

*Proof.* Let  $A_{>} := \{a \in A : a \geq \frac{\epsilon}{2}\}$  and  $A_{<} := A - A_{>}$ . Therefore

$$\begin{aligned} n\epsilon &\leq \sum_{a \in A} a = \sum_{a \in A_{<}} a + \sum_{a \in A_{>}} a \\ &\leq \sum_{a \in A_{<}} \frac{\epsilon}{2} + \sum_{a \in A_{>}} 1 \\ &= |A_{<}| \frac{\epsilon}{2} + |A_{>}| \end{aligned}$$

By rearranging and algebra,  $|\{a \in A : a \geq \frac{\epsilon}{2}\}| \equiv |A_{>}| > n\frac{\epsilon}{2}$  as required.  $\square$

*Proof of Lemma 13.* First,

$$\prod_{i=1}^{\infty} \left[1 - \frac{\alpha_i}{i}\right] \leq \exp \left[ - \sum_{i=1}^{\infty} \frac{\alpha_i}{i} \right] \quad (25)$$

Equation (25) follows since  $1 - a \leq \exp(-a)$  for all  $a$ .

Now since  $\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n a_n = \epsilon$ , we have for any  $N$  there exists an  $n > N$  such that  $\frac{1}{n} \sum_{i=1}^n a_n > \frac{\epsilon}{2}$ . Let  $n_1 = 0$  then inductively choose  $n_i = \min \left\{ n : n > \frac{8(n_{i-1}+1)}{\epsilon} \wedge \frac{1}{n} \sum_{i=1}^n a_i > \frac{\epsilon}{2} \right\}$

By Lemma 17,

$$\left| \left\{ i \leq n_j : a_i \geq \frac{\epsilon}{4} \right\} \right| \geq n_j \frac{\epsilon}{4} \quad (26)$$

Therefore

$$\sum_{i=n_j+1}^{n_{j+1}} \frac{\alpha_i}{i} \geq \sum_{i=n_{j+1}-\frac{\epsilon}{4}n_{j+1}+n_j+1}^{n_{j+1}} \frac{1}{n_{j+1}} \quad (27)$$

$$\geq \sum_{i=(1-\frac{\epsilon}{8})n_{j+1}}^{n_{j+1}} \frac{1}{n_{j+1}} = \frac{\epsilon}{8} \quad (28)$$

Equation (27) follows from (26) and because  $\frac{1}{i}$  is a decreasing function. (28) follows from the definition of  $n_j$  and algebra. Therefore

$$\begin{aligned} \sum_{i=1}^{\infty} \frac{\alpha_i}{i} &= \lim_{k \rightarrow \infty} \sum_{j=1}^k \sum_{i=n_j+1}^{n_{j+1}} \frac{\alpha_i}{i} \\ &\geq \lim_{k \rightarrow \infty} \sum_{j=1}^k \frac{\epsilon}{8} = \infty \end{aligned} \quad (29)$$

Finally, substituting Equation (29) into (25) gives,

$$\prod_{i=1}^{\infty} \left[1 - \frac{\alpha_i}{i}\right] = 0$$

as required.  $\square$

## B Table of Notation

Symbol	Description
$\mathcal{Y}$	Set of possible actions
$\mathcal{O}$	Set of possible observations
$\mathcal{R}$	Set of possible rewards
$\mu, \nu$	Environments
$y$	An action.
$x$	An observation/reward pair
$r$	A reward
$o$	An observation
$\llbracket expr \rrbracket$	The delta function. $\llbracket expression \rrbracket = 1$ if <i>expression</i> is true and 0 otherwise.
$\neg b$	The <i>not</i> function. $\neg 0 = 1$ and $\neg 1 = 0$ .
$\pi$	A policy.
$\chi$	An infinite binary string. $\chi_k = 1$ if an agent starts exploring at time-step $k$ .
$\bar{\chi}$	An infinite binary string. $\bar{\chi}_k = 1$ if an agent is exploring at time-step $k$ .
$\dot{\chi}^h$	An infinite binary string. $\dot{\chi}_k^h = 0$ if an agent will not explore for the next $h$ time-steps.
$\alpha$	An infinite binary string.
$\psi$	An infinite random binary string sampled from the coin flip measure.
$t, n, i, j, k$	Time indices.
$yx_{<t}$	A sequence of action/observation/reward sequences. Splits into $y_1 o_1 r_1 y_2 o_2 r_2 \cdots y_{t-1} o_{t-1} r_{t-1}$ .
$yx_{<t}^{\mu, \pi}$	The sequence of action/reward/observations seen in deterministic environment $\mu$ when playing policy $\pi$ .
$V_{\mu}^{\pi}(yx_{<t})$	The value of playing policy $\pi$ in environment $\mu$ starting at history $yx_{<t}$ .
$V_{\mu}^*(yx_{<t})$	The value of playing the optimum policy $\pi$ in environment $\mu$ starting at history $yx_{<t}$ .
$\pi_{\mu}^*$	The optimum policy in environment $\mu$ .
$H_t(p)$	The $p$ -percentile horizon.
$0^{h+1}$	A binary string consisting of $h + 1$ zeros.